COHORT-BASED LEARNING FROM USER EDITS

BACKGROUND

The continued proliferation of digital content items has led to an increase in the availability of such content items, as well as an increase in the availability of electronic devices and applications used for consuming these content items. For instance, users read text-based content items, such as electronic books (eBooks), magazines, newspapers, documents, or other textual content on an assortment of electronic devices. Many of these text-based content items were originally created as physical printed items. Thus, to make these printed items available as digital content items, the text may be scanned or otherwise imaged, and then the text may be recognized using automated text recognition technology, such as optical character recognition. However, automated text recognition techniques may be susceptible to errors. These errors can require significant manual corrections before the content item can be finalized for consumer use.

BRIEF DESCRIPTION OF THE DRAWINGS

The detailed description is set forth with reference to the accompanying figures. In the figures, the left-most digit of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical items or features.

FIG. **1** illustrates an example architecture that may implement a platform for assisting users in generating finalized works.

FIG. **2** illustrates details of an example computing device associated with the architecture of FIG. **1**.

FIG. **3** illustrates details of an example service provider associated with the architecture of FIG. **1**.

FIG. **4** illustrates example text images and corresponding edits or corrections.

FIG. **5** illustrates an example process for generating a character recognition-based work and modifying a global data set.

DETAILED DESCRIPTION

This disclosure describes, in part, a platform for assisting users in generating finalized works based on previous works generated using automated text recognition technology, such as optical character recognition. This platform includes applications that receive input from a variety of different sources including printed items, digital items, and the like. For example, text, charts, graphs, figures, and/or other content may be scanned or otherwise imaged, and a resulting digital file may be utilized as an input to one or more applications of the present disclosure. Additionally, existing tif, gif, pdf, doc, and/or other like electronic files may also be utilized as inputs to such applications. An exemplary application may combine such inputs into a single rough work. However, the rough work may contain multiple errors. For example, while such a rough work may include content from the inputs, such content may be improperly located on an example page of the work, may have poor image quality, may be improperly formatted or aligned, and/or made be otherwise unacceptable for processing by known automated text recognition technologies.

As a result, a user may utilize the application to revise the rough work by performing one or more manual corrections thereto. Such corrections may include, for example, aligning text, figures, charts, and/or other graphics contained in the rough work, organizing the text into appropriate paragraphs, enhancing the quality of one or more images contained in the rough work, and the like. Such manual corrections by the user may result in a relatively clean work suitable for processing by the automated text recognition technology employed by the application.

For example, one or more of the platforms described herein may include a recognition module configured to process such clean works. As part of such processing, the recognition module may analyze the clean work using text recognition technology. The recognition module may output and/or otherwise generate a first character recognition-based work including edits made automatically by the recognition module. As will be described with respect to the various embodiments discussed below, such automatically-made "edits" may comprise letters, numbers, symbols, and/or other characters that are automatically and/or semi-automatically changed by the recognition modules and/or other modules described herein. For each automatically-made edit, the recognition module may also output respective information characterizing the edit such that the edit may be categorized and/or easily accessed for use by the modules described herein in a further editing process. For example, the recognition module may output, for each automatically-made edit, a respective Unicode corresponding to the character changed by the recognition module, a metric indicative of the bounds of the changed character relative to, for example, a text image of the work, a confidence score associated with the edit, and/or one or more alternate suggested edits associated with the automatically-made edit. Further, the recognition module may also output a respective confidence score corresponding to each alternate suggested edit.

One or more of the platforms described herein may also include a processing module configured to post-process one or more outputs of the recognition module, and to thereby generate a processed work. For example, upon receiving the character recognition-based work and/or other outputs from the recognition module, the processing module may identify at least one of the automatically-made edits as being of questionable accuracy. For example, the processing module may compare the confidence score of each respective automatically-made edit to a confidence score threshold. In such an embodiment, the processing module may characterize automatically-made edits having a confidence score less than the confidence score threshold as being of questionable accuracy, and may characterize automatically-made edits having a confidence score greater than the confidence score threshold as being of acceptable accuracy.

Additionally, upon receiving outputs from the recognition module, the processing module may extract and/or otherwise determine a unique character signature indicative of and/or otherwise associated with an edit automatically-made by the recognition module. Such character signatures may include, for example, a shape identifier, a boundary identifier, a location identifier, and/or other like identifiers indicative of a character of the automatically-made edit. For example, shape context features, and/or other like features known in the art configured to uniquely identify one or more contours of the various characters associated with the automatically-made edit may be extracted by the processing module during generation of the processed work. The various outputs of the recognition module and of the processing module may be stored in one or more data stores, such as within a works data store associated with a local computing device, for future use.

One or more of the platforms described herein may also include a manual correction module configured to receive manual corrections made to the various automatically-made